Towards Automatic ICD Coding via Knowledge Enhanced Multi-Task Learning

Xinhang Li
Department of Computer Science and Techonology,
Tsinghua Univerisity
Beijing, China
xh-li20@mails.tsinghua.edu.cn

Xiangyu Zhao*
School of Data Science, City Univerisity of Hong Kong
Hong Kong
xianzhao@cityu.edu.hk

Yong Zhang*
Department of Computer Science and Technology,
Tsinghua Univerisity
Beijing, China
zhangyong05@tsinghua.edu.cn

Chunxiao Xing
Department of Computer Science and Technology,
Tsinghua Univerisity
Beijing, China
xingcx@tsinghua.edu.cn

ABSTRACT

The aim of ICD coding is to assign International Classification of Diseases (ICD) codes to unstructured clinical notes or discharge summaries. Numerous methods have been proposed for automatic ICD coding in an effort to reduce human labor and errors. However, existing works disregard the data imbalance problem of clinical notes. In addition, the noisy clinical note issue has not been thoroughly investigated. To address such issues, we propose a knowledge enhanced Graph Attention Network (GAT) under multi-task learning setting. Specifically, multi-level information transitions and interactions have been implemented. On the one hand, a large heterogeneous text graph is constructed to capture both intra- and inter-note correlations between various semantic concepts, thereby alleviating the data imbalance issue. On the other hand, two auxiliary healthcare tasks have been proposed to facilitate the sharing of information across tasks. Moreover, to tackle the issue of noisy clinical notes, we propose to utilize the rich structured knowledge facts and information provided by medical domain knowledge, thereby encouraging the model to focus on the clinical notes' noteworthy portion and valuable information. The experimental results on the widely-used medical dataset, MIMIC-III, demonstrate the advantages of our proposed framework.

CCS CONCEPTS

• **Information systems** → **Document representation**; • **Computing methodologies** → **Information extraction**; • **Applied computing** → *Health informatics.*

KEYWORDS

ICD coding, multi-task learning, knowledge graph

*Xiangyu Zhao and Yong Zhang are the corresponding authors.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3615087

ACM Reference Format:
Xinhang Li, Xiangyu Zhao, Yong Zhang, and Chunxiao Xing. 2023. Towards Automatic ICD Coding via Knowledge Enhanced Multi-Task Learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3583780.3615087

1 INTRODUCTION

ICD coding is a standardized method for extracting diagnosis and procedure codes from clinical texts regarding a patient encounter [5]. Due to the importance of ICD codes, which are widely used for clinical both research and healthcare purposes, it has garnered considerable attention [39, 46, 60–62]. As manual code assignment is labor-intensive and error-prone, automatic ICD coding from unstructured texts has been proposed and investigated by the research community [47, 48].

Generally, existing automatic ICD coding methods can be categorized as either traditional feature extraction [21, 32] or deep learning-based approaches [6, 23, 27, 29, 33, 49, 58, 63]. Among others, convolutional methods [23, 27, 49, 63] are the most widely-used encoding framework and outperform alternatives. However, considering the long-tail issue of code frequency, several methods have recently attempted to enhance ICD coding by incorporating external information such as ICD codes [44, 63], code hierarchy and code co-occurrence [6, 42, 63]. Although efficient, these methods only address the long-tail problem from the perspective of label space, ignoring the inherent imbalance problem of data itself.

Nonetheless, the data imbalance problem is the more intrinsic and fundamental issue and is also the cause of the long-tail problem of code frequency. In particular, the distribution of semantic concepts in clinical notes is usually unbalanced. Some patient symptoms (*e.g.*, cough) are extremely common, whereas other symptoms (*e.g.* loss of smell) are extremely uncommon in clinical notes. Besides, the frequency of various types of medicine exhibits a long-tailed distribution as well. For example, anti-fever medications are more prevalent than anti-Alzheimer drugs. This imbalance of data leads directly to the long tail distribution of the ICD code, where the top 10% of high-frequency codes in MIMIC-III accounted for 85% of total occurrence [63]. Neural ICD coding methods trained with

these unbalanced data would not learn extensive semantic information, particularly for concepts representing symptoms, medications, treatment procedures, etc, related to the tail ICD codes.

Inherent relationships among medical concepts and clinical notes are often utilized to compensate for the imbalance-induced deficiency. As an example, *aspirin* is a common pain reliever usually prescribed for toothache and headache. It is also used for fever diseases like colds and flu. However, when combined with *statins*, it is typically used to treat *atherosclerotic cardiovascular disease* (ASCVD), which occurs less frequently than headaches or flu. Therefore, if a model could discover correlations between *aspirin* and *statins*, the prediction space would be narrowed to a subset of diseases associated with ASCVD. In addition, there are relations between different clinical notes representing different patients: if different clinical notes have similar symptoms, treatments, etc., they are likely to share similar ICD codes. Thus, how to learn the semantics of clinical notes, with considering multi-level correlations, is a crucial and challenging task. Besides, clinical notes also exhibit a problem with noisy text that is difficult to read. Specifically, they always include multiple lengthy textual narratives (more than 1500 tokens on average), whereas only a small portion of tokens are relevant for specific ICD codes. In other words, an abundance of information in clinical notes is redundant or misleading for ICD coding task [63]. Therefore, capturing the informative portion of lengthy clinical notes is an additional challenge.

In light of above issues, we propose a Knowledge Enhanced GAT with Multi-Task Learning (KEMTL) for automatic ICD coding. Firstly, to capture the informative portion of lengthy clinical notes, we employ medical domain knowledge to guide the contextual feature extraction by leveraging structured entity (medical concepts) and relation information. To discover medical concepts contained in clinical notes, we employ a medical domain knowledge system called Unified Medical Language System (UMLS) [1]. Then, multiple levels of information transition and interaction will be utilized to address the data imbalance problem [24]. To model the information correlations within both intra- and inter-notes, we construct a large heterogeneous text graph containing two types of nodes: concept nodes and document nodes by aligning the knowledge of entities and relationships in UMLS to texts in the corpus. Given such a text graph, we formulate the clinical note modeling process as the document node embedding learning process in the graph. We employ a Graph Attention Network (GAT) [43] on the text graph to enable both local contextual feature encoding and inter-note global feature encoding by gathering high-order neighbor information. In this way, the co-occurrence and semantics of medical concepts, as well as the relationships between different clinical notes, can be utilized extensively to enhance imbalanced text encoding.

On the other hand, we find that various healthcare tasks (*e.g.*, treatment recommendation and mortality prediction) are related to the ICD coding task and should not be considered independent. For instance, the ICD codes representing the diseases would limit the treatment procedures that can be used; knowledge of diseases and treatment procedures are the triggers for mortality outcomes, and vice versa. Therefore, we present a multi-task learning framework to model them jointly, aiming at across-task information sharingk [26, 45, 59] to benefit ICD coding. To enable MTL, we devise two information-sharing mechanisms between distinct tasks. Firstly, we propose a widely-used MTL architecture with a global shared layer in which multiple tasks share the same GAT encoding layer for text representation. Through the global shared encoding layer, each individual task can thus gain access to the shared knowledge of all other tasks. After that, each task has its own distinct classification layer. Secondly, considering the need to maintain the uniqueness of each clinical task, we further propose a second MTL architecture, in which each task has its own task-specific GAT encoding layer and a global shared layer is designed for all tasks to facilitate information sharing.

Contributions of this paper are summarized as following:

- We propose a knowledge enhanced GAT with multi-task learning (KEMTL) for automatic ICD coding. Compared to previous studies, KEMTL can utilize multi-level information transition and interaction to improve the overall performance of ICD coding.
- We introduce medical domain knowledge to improve the local contextual feature extraction of a single note, and utilize GAT model to facilitate information transfer among and within the notes by constructing an entity-level text graph for the entire corpus.
- We design two auxiliary healthcare tasks and propose two information-sharing mechanisms to utilize knowledge from multiple tasks, considering both the task specificity and the information sharing across tasks.
- Extensive experiments are conducted, and the results demonstrate that KEMTL achieves state-of-the-art performance on the widely used MIMIC-III medical datasets.

2 PRELIMINARY

In this paper, we use the GAT [43] to encode the text graph constructed from the clinical notes and medical domain knowledge. The advantage of GAT is that it leverages attention mechanism to consider different weights from neighbor nodes, which enables the model to concentrate on important adjacent nodes. Specifically, the GAT layer aggregates features of each node in the graph as well as its one-hop neighbors as new features. The detailed process on the t^{th} layer for node v is formalized as Equation (1) - (2).

$$\boldsymbol{h}_v^{(t)} = \sigma\Big(\sum_{u \in \mathcal{N}(v) \cup \{v\}} \alpha_{vu} \mathbf{W}^{(t)} \boldsymbol{h}_u^{(t-1)}\Big) \tag{1}$$

$$\begin{aligned} \alpha_{vu} &= \mathrm{softmax}\big(f(\boldsymbol{a}_t^T [\mathbf{W}^{(t)} \boldsymbol{h}_v^{(t-1)} || \mathbf{W}^{(t)} \boldsymbol{h}_u^{(t-1)}])\big) \\ &= \frac{\exp\big(f(\boldsymbol{a}_t^T [\mathbf{W}^{(t)} \boldsymbol{h}_v^{(t-1)} || \mathbf{W}^{(t)} \boldsymbol{h}_u^{(t-1)}])\big)}{\sum_{j \in \mathcal{N}(v) \cup \{v\}} \exp\big(f(\boldsymbol{a}_t^T [\mathbf{W}^{(t)} \boldsymbol{h}_v^{(t-1)} || \mathbf{W}^{(t)} \boldsymbol{h}_j^{(t-1)}])\big)} \end{aligned} \tag{2}$$

where $\mathbf{W}^{(t)}$ is the weight matrix, α_{vu} is the attention coefficient of node u to v, $\mathcal{N}(v)$ denotes the neighbors of node v, f donates the *LeakyReLU* function and $\boldsymbol{a}_t$ is the weight vector.

3 METHODOLOGY

3.1 Overall Architecture

As shown in Figure 1, our proposed KEMTL consists of two components: knowledge enhanced text graph constructor and multi-task

[1] https://www.nlm.nih.gov/research/umls/index.html

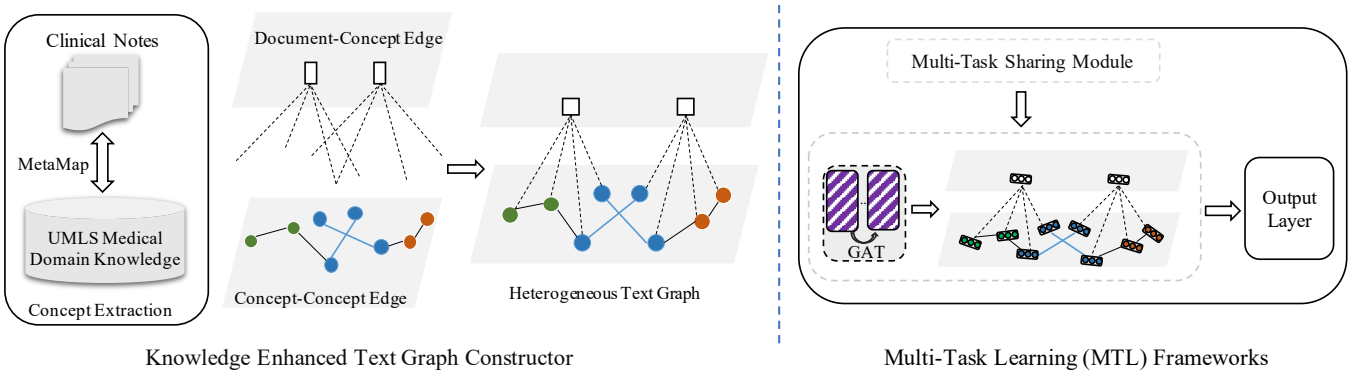

Figure 1: Overall Architecture of KEMTL.

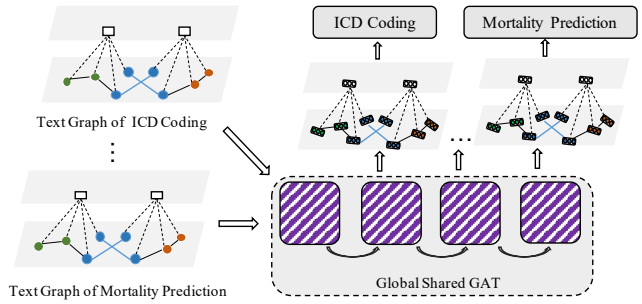

Figure 2: Model Structure of KEMTL-uni.

learning frameworks. The text graph constructor transforms the clinical notes into the entity-level text graph leveraging medical domain knowledge for every single task. Then, nodes in each obtained text graph are encoded into low-dimensional vectors via GAT models.

Next, we propose a MTL framework to enhance the node representations with across-task information. Given K related tasks $\{T_k\}_{k=1}^{K}$, MTL aims to enhance each task T_k by utilizing the knowledge contained in all tasks [57]. Usually a task T_k is accompanied by a training dataset D_k containing N_k samples denoted as:

$$D_k = \{(X_i^{(k)}, Y_i^{(k)})\}_{i=1}^{N_k} \tag{3}$$

where $X_i^{(k)}$ and $Y_i^{(k)}$ are i^{th} training instance and its label in the k^{th} task respectively. The key factor of multi-task learning is the sharing scheme, which determines the form of knowledge sharing among all tasks. In this paper, we devise two multi-task learning frameworks with different sharing schemes. Finally, each clinical note is represented as a vector and the vector is fed to the task-specific output layer to make the final prediction for each task.

3.2 Knowledge Enhanced Text Graph Constructor

We use the well-known UMLS to help build an entity-level heterogeneous text graph for each task. The graph involves two types of nodes: *concept node* and *document node*. For the task k, supposing there are $\mathbb{N}_k$ clinical notes, then the number of nodes $|V_k|$ in the generated text graph G_k is the number of clinical notes $\mathbb{N}_k$ plus the number of unique concepts extracted from its MetaMap [2] [3].

Generally speaking, the text graph construction process consists of three steps: extracting medical concepts, identifying document-concept edges and identifying concept-concept edges. First of all, we feed each clinical note of task k into its MetaMap to obtain the medical-related concepts in it. In this way, we can better discover informative variable-length n-gram features and deep semantics of texts with the help of domain knowledge. Then, inspired by the previous study [53], we utilize the Term Frequency-Inverse Document Frequency (TF-IDF) of concepts in each clinical text to help identify the edges among concept and document nodes. Compared with other ways, *e.g.*, bag of words or term frequency only, to describe the document-concept relations, TF-IDF can better assess the importance of a concept to one of the documents within the whole corpus since it takes document frequency into consideration. We keep the top 10,000 concepts according to the TF-IDF values in the final text graph and leave out others. Finally, to identify concept-concept edges, we first initialize the edges according to relationship table of UMLS. Then, we further expand the edges based on the co-occurrence statistics in the corpus to utilize global concept co-occurrence information. In this way, both intra- and inter-note correlated medical concepts can be linked in the text graph. Specifically, we employ the popular measure for term associations Point-wise Mutual Information (PMI) to determine relations between two concept nodes. Formally, the PMI value between concept node i and concept node j is defined as:

$$PMI(i,j) = log\frac{\#W(i,j)\#W}{\#W(i)\#W(j)} \tag{4}$$

where $\#W(i)$ is the number of sliding windows in a corpus that contain concept i; $\#W(i,j)$ is the number of sliding windows in a corpus that contain both concept i and concept j; and $\#W$ is the total number of sliding windows in the corpus. In this paper, we set the window size to that of a clinical text to guarantee that the total number of sliding windows is the corpus size. A positive PMI value implies a high semantic correlation of concepts in a corpus, while a negative one indicates little or no semantic correlation in

[2] A tool for recognizing UMLS concepts in text: https://metamap.nlm.nih.gov/

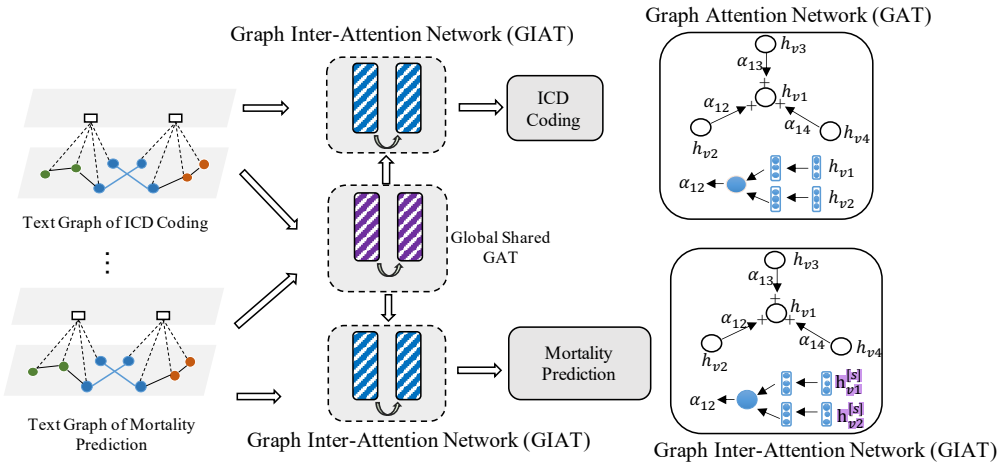

Figure 3: Model Structure of KEMTL-spec.

the corpus. Therefore, we add edges between concept pairs only when the PMI values are positive.

3.3 Multi-Task Learning (MTL) Frameworks

We use the GAT model to encode the above text graphs. On the basis of it, we then propose our MTL frameworks to improve the document-node representations using cross-task information. The intuition is that different healthcare tasks share some common information and knowledge, which can help improve the clinical note representation of the ICD coding task and contribute to better performance. In order to benefit from both the task specificity and the information sharing, we propose two models KEMTL-uni and KEMTL-spec by designing different ways to share the cross-task information.

3.3.1 Universal Sharing Hierarchical MTL (KEMTL-uni). Figure 2 shows the universal information sharing model KEMTL-uni. Different tasks share the same GAT encoding layer for text graph modeling. Besides, to mirror the inter-dependencies (*e.g.*, ICD coding → treatment recommendation → mortality prediction) of these tasks, we introduce a hierarchy between the tasks so that low-level tasks are supervised at lower levels of the GAT while keeping more complex interactions at deeper layers. After that, each task has its own specific layer for final prediction. As follows, for the task k, supposing its corresponding output layer is the t^{th} layer. Then, the convolution process is formalized as:

$$\boldsymbol{h}_v^{[k](t)} = GAT(\{\boldsymbol{h}_i^{[k](t-1)}\}_{i\in\mathcal{N}(v)\cup\{v\}}, \Theta^{(s)}) \tag{5}$$

Here, we use $GAT(\cdot,\cdot)$ as a shorthand for the graph attention network shown in Equation (1) to (2) and $\Theta^{(s)}$ refers to the parameters of GAT, which are shared by all three tasks in KEMTL-uni. In this way, every node of a single task can benefit from common knowledge and extra information from all other tasks. Besides, each task can make use of such information to receive different weights from neighbor nodes, which helps filter out noises and focus on important adjacent nodes.

3.3.2 Task Specific Sharing MTL (KEMTL-spec). Although KEMTL-uni is able to utilize global information from other tasks, it fails to model the task specificity and information-sharing simultaneously. For example, in the ICD coding task, disease-related nodes need to be emphasized and assigned high attention during the graph modeling process; whereas physical condition and instrumental physical indicators-related nodes should be valued in the mortality prediction task. To resolve this issue, we further propose KEMTL-spec with another information sharing scheme whose basic idea is to preserve the task specificity for each single task while sharing common knowledge among different tasks. The architecture of KEMTL-spec is shown in Figure 3. Each task has its own task-specific encoding layer in KEMTL-spec. Then, a global shared layer with another GAT is employed to capture the global information among all tasks. In order to make full use of information of different tasks, we propose a novel graph inter-attention network (GIAT) for each task. The intuition is that as external knowledge, information from the global shared component can provide common and task-invariant knowledge, which can help disambiguate the semantics of clinical concepts of every single task. Specifically, for t^{th} layer output of node v in graph G_k for task k, we first compute its task specific representation according to $(t-1)^{th}$ layer output:

$$\tilde{\boldsymbol{h}}_v^{[k](t)} = \sigma\Big(\sum_{u\in\mathcal{N}(v)\cup\{v\}} \alpha_{vu}\mathbf{W}^{[k](t)}\boldsymbol{h}_u^{[k](t-1)}\Big) \tag{6}$$

$$\alpha_{vu} = softmax\big(f(\boldsymbol{a}_t^T[\mathbf{W}^{[k](t)}\boldsymbol{h}_v^{[s](t-1)} || \mathbf{W}^{[k](t)}\boldsymbol{h}_u^{[s](t-1)}])\big) \tag{7}$$

where $\mathbf{W}^{[k](t)}$ is the weight matrix of task k, $\boldsymbol{h}_u^{[s](t-1)}$ and $\boldsymbol{h}_v^{[s](t-1)}$ are the output of the global shared $(t-1)^{th}$ layer. Note that in GIAT, the attention coefficient α_{vu} of node u to v is computed as the inner product of $\boldsymbol{h}_u^{[s](t-1)}$ and $\boldsymbol{h}_v^{[s](t-1)}$ rather than its task specific output followed by a softmax layer.

We compute the global output of the t^{th} layer based on the global shared GAT model:

$$\boldsymbol{h}_v^{[s](t)} = GAT(\{\boldsymbol{h}_i^{[s](t-1)}\}_{i\in\mathcal{N}(v)\cup\{v\}}, \Theta^{(s)}) \tag{8}$$

Algorithm 1: KEMTL training process.

1 **Initialization**:A set of models $\{\mathcal{M}_k\}_{k=1}^{K}$ for different tasks.
// Construct knowledge enhanced text graphs
2 **for** $k \leftarrow 1$ **to** K *tasks* **do**
3 Extract medical concepts of all the clinical notes in task k using MetaMap;
4 Identify document-concept edges by filtering unnecessary edges with TF-IDF;
5 Identify concept-concept edges using UMLS;
6 Expand concept-concept edges based on positive PMI as Equation (4);
7 Formulate the text graph G_k of task k with document nodes, concept nodes, document-concept edges and concept-concept edges;
8 **end**
// Multi-task learning
9 **for** $t \leftarrow 1$ **to** T *epochs* **do**
10 **for** $k \leftarrow 1$ **to** K *tasks* **do**
11 Obtain latent node embeddings of documents in task k via Equation (5) or Equation (9);
12 Compute the final output of task k via Equation (10);
13 **end**
14 Compute the overall loss $L(\Theta)$ from all K tasks according to the task-specific dataset D_k in Equation (3) and the task weights β_k via Equation (11);
15 **end**

where $\Theta^{(s)}$ represents the parameters of the global shared layer. Finally, the output of task k is the weighted sum of that of the task specific layer and global shared layer:

$$\boldsymbol{h}_v^{[k](t)} = \sigma(g^{k\to k}) \odot \tilde{\boldsymbol{h}}_v^{[k](t)} + \sigma(g^{s\to k}) \odot \boldsymbol{h}_v^{[s](t)} \tag{9}$$

where $g^{i\to k}(i \in s, k)$ controls the portion of information flow from the global shared layer and task-specific layer to task k respectively and will be learned during the training process. In this way, each task can not only take the advantage of the common knowledge through the global shared GAT but keep its own identity with the help of the task-specific layer.

3.4 Output Layer and Training

After node encoding, each clinical note is represented as a vector $\boldsymbol{d}$ and each task has its specific layer for final prediction. Following previous studies [29, 49], we regard healthcare tasks as multi-label classification problem. The final output layer is formulated as:

$$\hat{\boldsymbol{y}} = f(\mathbf{W}_y \boldsymbol{d} + \boldsymbol{b}_y) \tag{10}$$

where $\mathbf{W}_y$ and $\boldsymbol{b}_y$ denote the weight matrix and the bias vector. Here we use *sigmoid* as the activation function f.

The objective of the training process is to minimize the cross-entropy of the predicted and true distributions in all tasks:

$$L(\Theta) = -\sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{j=1}^{C_k} \beta_k y_{ij}^{(k)} log(\hat{y}_{ij}^{(k)}) \tag{11}$$

where N_k, C_k and β_k refer to the number of training samples, the number of classes and the loss weight of task k, respectively. $y_{ij}^{(k)}$ is the ground-truth label, $\hat{y}_{ij}^{(k)}$ is the predicted probability and Θ is the set of all trainable parameters. Specifically, our proposed framework takes one task as the main task each time, which has a large loss weight. The other two tasks, which have small loss weights, are treated as auxiliary tasks to provide knowledge for more robust and accurate prediction of the main task.

To better understand our proposed KEMTL, we describe the whole training process in Algorithm 1, which consists of two disentangled parts. First, the knowledge enhanced text graphs for all the tasks are constructed by extracting medical concepts and identifying the edges (line 2-8). Then, we obtain the task-specific outputs for each task through the task-specific GAT/GIAT (line 9-13). Finally, the overall loss $L(\Theta)$ is computed by integrating the loss from all the tasks with task weights for achieving multi-task learning (line 14-15).

4 EXPERIMENTS

4.1 Experiment Setup

4.1.1 Datasets. In order to fully evaluate the effectiveness of our proposed method, we perform experiments on the publicly available MIMIC-III [16] event note which is an open-access collection of datasets and used as the benchmarking datasets by previous work on ICD coding [6, 29, 49, 63]. There are different event notes in MIMIC-III for each stay of patients in the ICU, including discharge summary reports, nursing notes, radiology notes etc.

Specifically, for ICD coding, we use discharge summaries that are tagged manually with a set of ICD-9 codes. The experiments are performed with two different settings to formulate two datasets: 1) MIMIC-III Full dataset: we use all discharge summaries with all 8,921 labels as the corpus; 2) MIMIC-III 50 dataset: we only predict 50 most frequent labels, and filter each split in MIMIC-III Full dataset down to the instances that have at least one of the top 50 codes. As for the two auxiliary tasks, we focus on predicting the top 50 most frequent labels using discharge summaries for treatment recommendation [22] and we use the nursing notes for mortality prediction as previous works [14, 25] did. Due to the limited space, the descriptions of datasets are shown in Table 1.

4.1.2 Baseline Methods. To verify the effectiveness, we compare the proposed KEMTL model with several state-of-the-art methods for ICD coding and the two auxiliary tasks (treatment recommendation and mortality prediction).

For automatic ICD coding, we choose:

- CAML (DR-CAML) [29] first introduces CNN to encode clinical notes and get predictions via dot product between text and code embeddings.
- MSATT-KG [49] utilizes multi-scale feature attention and GCN to capture the relations between ICD codes.
- HyperCore [6] converts the embeddings of ICD codes into hyperbolic space and then employs GCN on the co-occurrence graph to model the hierarchical information.
- MultiResCNN [23] utilizes multi-filter residual CNN for encoding medical texts.

Table 1: Statistics of MIMIC-III Datasets.

Task	Category	Split	#Texts	#Labels
ICD Coding (Top 50)	Discharge Summary	Train	8,067	50
		Val	1,574	
		Test	1,730	
ICD Coding (Full)	Discharge Summary	Train	47,719	8,921
		Val	1,631	
		Test	3,372	
Treatment Recommendation	Discharge Summary	Train	6,884	50
		Val	983	
		Test	1,967	
Mortality Prediction	Nursing Note	Train	1,632	2
		Val	233	
		Test	466	

- LAAT (JointLAAT) [44] proposes a hierarchical joint learning mechanism to alleviate the imbalance problem of labels.
- Fusion [27] designs complex feature compression and aggregation to better model the medical texts.
- ISD [63] leverages self-distillation mechanism to enhance the interactive shared representation networks for ICD coding.
- MSMN [54] considers the synonyms of codes in a graph matching manner to achieve more robust performance.

Note that the automatic ICD coding task is different from the medication recommendation task although they both aim at predicting a set of medical codes. While the automatic ICD coding task predicts ICD codes using medical texts, the medication recommendation task leverages the diagnoses and procedure codes to predict medication codes. Thus, the approaches for medication recommendation task, such as RETAIN [11], LEAP [56], GAMENet [37] and SafeDrug [50] are orthogonal to our KEMTL and are excluded for comparison.

For treatment recommendation, since existing treatment recommendation approaches are mainly based on structured features rather than unstructured clinical texts, we take some famous text classification approaches instead:

- Bi-GRU [10] utilizes bidirectional GRU to encode medical texts with a single feedforward layer for prediction.
- Text-CNN [18] applies CNN on concatenated word embeddings for text classification.
- HAN [52] employs hierarchical attention to capture the multi-level information in texts.

For mortality prediction, we have:

- H_CNN [14] leverages hierarchical embeddings from sentence-level to document-level for document classification.
- DKGAM [7] introduces external knowledge from the word-level to help text classification and we replace the words with concepts in our settings.
- AK-DNN [25] incorporates medical domain knowledge to enhance the text representations for prediction.

4.1.3 Implementation Details. The experiments are implemented on the server with an Intel Xeon E5-2640 CPU, a 188GB RAM and four NVIDIA GeForce RTX 2080Ti GPUs. All the models are implemented using PyTorch 1.6.0. The model parameters are optimized using Adam [19] optimizer. For KEMTL-uni, we use a four-layer GAT for text graph encoding and the output layer for ICD coding is the 2nd layer. For KEMTL-spec, we use a two-layer GAT/GIAT for text graph encoding and information sharing that allows message passing among nodes two hops away. As a result, although there is no direct document-document edge in the graph, the multi-layer GAT is still capable to learn the inter-text correlations and enables the information exchange between pairs of notes. For baseline methods, we directly use the results reported in the original papers for ICD coding since all the approaches use the same settings. For Text-CNN and HAN, we use the source code provided by the authors and tune the parameters according to the instructions provided by them.

4.2 Overall Performance

We report the experimental results of our KEMTL in Table 2. For fair comparison with previous studies, we employ a variety of evaluation metrics following their routine, including micro F_1, macro F_1, Area Under the ROC Curve (AUC) and Precision at n. From the results, we can see that the model performance is gradually improved with the increase of interaction between clinical notes and ICD codes. In CNN-based methods, including CAML, DR-CAML, MultiResCNN, HyperCore and MSATT-KG, MSATT-KG achieves the best performance and a significant improvement over previous works due to its strong ability to jointly capture note-note and code-code correlations with the help of attention mechanism and GNN. Following it, more recent approaches have focused on designing effective mechanisms to model both intra- and inter-note correlations and the improvements over MSATT-KG indicate the effectiveness of these approaches.

However, the improvements brought by optimization of the model structure are marginal while our proposed KEMTL breaks through the bottleneck by effective knowledge enhancement and multi-task learning frameworks. We can observe that KEMTL-uni and KEMTL-spec obtain promising performance on both two datasets, demonstrating the effectiveness and robustness of KEMTL.

More specifically, on MIMIC-III 50, KEMTL-spec shows an improvement of macro AUC 1.7%, macro F_1 1.2% and p@5 2.9% to Fusion, the most competitive model. Besides, our KEMTL-uni obtains the highest score of 95.5% on micro AUC value. These convincing results indicate the superiority of introducing knowledge from different related tasks compared with the complex feature compression and aggregation approaches used in Fusion, and thus clearly demonstrate the effectiveness of our proposed multi-task learning framework on ICD coding.

Moreover, in terms of MIMIC-III Full dataset, MSMN is the strongest baseline since it considers the synonyms of codes in a graph matching manner to alleviate the long-tail and noisy text issues. Compared with it, our KEMTL can resolve these issues by integrating the structured entities and relations of medical domain knowledge into the informative local and global feature extraction process to capture both intra- and inter-note knowledge. The experimental results indicate the effectiveness of our method: KEMTL-spec has a 0.5% improvement on Macro AUC, 0.8% improvement on

Table 2: Evaluation Results of ICD Coding on MIMIC-III 50 and MIMIC-III Full Datasets.

Model	MIMIC-III 50					MIMIC-III Full				
	AUC		F1		P@5	AUC		F1		P@8
	Macro	Micro	Macro	Micro		Macro	Micro	Macro	Micro	
CAML	87.5	90.9	53.2	61.4	60.9	89.5	98.6	8.8	53.9	70.9
DR-CAML	88.4	91.6	57.6	63.3	61.8	89.7	98.5	8.6	52.9	69.0
MultiResCNN	89.9	92.8	60.6	67.0	64.1	91.0	98.6	8.5	55.2	73.4
HyperCore	89.5	92.9	60.9	66.3	63.2	93.0	98.9	9.0	55.1	72.2
MSATT-KG	91.4	93.6	63.8	68.4	64.4	91.0	99.2	9.0	55.3	72.8
LAAT	92.5	94.6	66.6	71.5	67.5	91.9	98.8	9.9	57.5	73.8
JointLAAT	92.5	94.6	66.1	71.6	67.1	92.1	98.8	10.7	57.5	73.5
Fusion	93.1	95.0	68.3	72.5	67.9	91.5	98.7	8.3	55.4	73.6
ISD	93.5	94.9	67.9	71.7	68.2	93.8	99.0	11.9	55.9	74.5
MSMN	92.8	94.7	68.3	72.5	68.0	95.0	99.2	10.3	**58.4**	75.2
KEMTL-uni	93.5	**95.5***	68.1	71.8	69.4	94.0	**99.6***	11.2	57.1	74.7
KEMTL-spec	**94.8***	94.2	**69.5***	**72.9***	**70.8***	**95.3***	99.4	**12.7***	58.3	**75.6***

"*" indicates the statistically significant improvements (i.e., two-sided t-test with $p < 0.05$) over the best baseline.

Table 3: Evaluation Results on MIMIC-III Dataset for Ablation Study.

Model	ICD Coding (Top 50)					Treatment Recommendation					Mortality Prediction
	AUC		F1		P@5	AUC		F1		P@5	AUC
	Macro	Micro	Macro	Micro		Macro	Micro	Macro	Micro		
GMTL-single	89.5	90.0	66.2	67.0	64.4	84.5	85.6	63.5	65.5	66.1	90.7
KEMTL-single	90.2	91.2	66.5	68.2	64.8	85.0	86.1	64.3	65.1	66.7	91.3
GMTL-uni	92.1	93.5	67.2	69.2	68.2	85.2	86.6	64.3	66.4	67.5	95.5
GMTL-spec	93.3	92.1	68.7	71.2	69.7	88.7	89.5	69.7	70.1	70.5	96.2

Macro F_1 score, and KEMTL-uni has 0.6% improvement on Micro AUC, 1.2% improvement on Micro F_1 score to ISD, respectively.

4.3 Ablation Study

To investigate the effectiveness of our proposed components of the method, we report the evaluation results of ablated versions of our KEMTL in Table 3. We proposed 4 methods by incrementally removing a component of the final model: GMTL-uni and GMTL-spec denote the word-level GAT-based universal sharing MTL and task-specific sharing MTL models without medical domain knowledge. In these two models, we build the text graph only based on the word-level co-occurrence and document-word relations as previous study [53] did. KEMTL-single refers to the single-task model with only knowledge enhanced GAT as encoders for each task. And GMTL-single is the word-level GAT-based single-task model without medical domain knowledge. From the table, we can see that GMTL-uni, GMTL-spec and KEMTL-single all perform better than the base model GMTL-single, confirming the benefits of both medical domain knowledge and MTL of our framework. Besides, multi-task models GMTL-uni and GMTL-spec have better performances than the single-task model with knowledge KEMTL-single, indicating the proposed auxiliary tasks are beneficial to the ICD coding task. Moreover, compared with some recent baselines such as MultiResCNN [23] and HyperCore [6], our base model GMTL-single has been able to achieve comparable or even better results. This shows the effectiveness to utilize both intra- and inter-note correlations between different medical concepts with text graph encoding methods.

4.4 Effectiveness on Auxiliary Tasks

To further demonstrate the effectiveness of our proposed KEMTL, we also report the performance on the two auxiliary tasks in Table 4 and Table 5 respectively. Specifically, for treatment recommendation and mortality prediction tasks, we also see pronounced improvements. One thing worthy to note is that our KEMTL models also obtain a high macro score while other models always have a lower one. This is because, for multi-label classification task, the macro-averaged metrics place more emphasis on relatively tail label prediction compared with micro-averaged values. Since there is insufficient training data for those rare labels, the performance of single-task models would definitely suffer from it. As our model can make use of global shared information from multiple tasks and multi-level information correlations, it can perform well on all labels. For mortality prediction task, AK-DNN performs best among all single-task learning methods, since it incorporates the medical domain knowledge to help enhance the text representation. As our KEMTL can not only utilize such medical knowledge to strengthen text representations, but also can benefit from the common information and knowledge of multiple tasks, it undoubtedly achieves the best results.

Table 4: Evaluation Results of Treatment Recommendation.

Model	AUC		F1		P@5
	Macro	Micro	Macro	Micro	
Bi-GRU	81.7	85.2	53.2	63.4	71.2
Text-CNN	82.1	85.4	59.5	65.6	70.4
HAN	82.9	86.1	60.1	66.0	70.2
KEMTL-uni	86.3	87.2	65.4	67.7	69.9
KEMTL-spec	**89.9**	**90.7**	**71.8**	**73.1**	**71.6**

Table 5: Evaluation Results of Mortality Prediction.

Model	H_CNN	DKGAM	AK-DNN	KEMTL-uni	KEMTL-spec
AUC	80.2	81.1	87.3	99.7	**99.8**

Table 6: Evaluation Results with Different Backbones.

Model	AUC		F1		P@5
	Macro	Micro	Macro	Micro	
GMTL-single GCN	88.7	89.6	62.7	64.2	62.9
KEMTL-spec GCN	92.0	91.8	68.9	70.0	69.4
GMTL-single GraphSAGE	89.3	90.2	65.6	67.8	64.5
KEMTL-spec GraphSAGE	90.8	91.2	66.4	68.7	65.3
GMTL-single GAT	89.5	90.0	66.2	67.0	64.4
KEMTL-spec GAT	**94.8**	**94.2**	**69.5**	**72.9**	**70.8**

4.5 Generalization with Different Backbones

Since our proposed KEMTL is a general multi-task learning framework that can be applied to any graph-based models, we conduct an experiment by replacing GAT with two widely-used graph embedding models, which are GCN [20] and GraphSAGE [15], to verify the generalization of KEMTL. Specifically, we compare the performance on ICD coding of variant without knowledge enhancement and multi-task learning (denoted by GMTL-single Backbone) and the full KEMTL-spec variant (denoted by KEMTL-spec Backbone) for each backbone model. As shown in Table 6, by introducing KEMTL as the training framework, the performance of each backbone model significantly improves, which further demonstrates the effectiveness and the generalization of KEMTL.

4.6 Sensitivity Analysis

In order to illustrate that our proposed multi-task learning framework is a good solution to the problem of insufficient training data, we conduct a sensitivity analysis on the size of data. Specifically, we vary the data sizes by randomly sampling different ratios of the training data for training and test them on the whole test sets of three tasks. Figure 4 show the experimental results of AUC scores of our KEMTL and the corresponding most competitive baseline models on ICD coding. From the figure, we can readily see that KEMTL-uni and KEMTL-spec consistently outperform the baseline model. Besides, the performance gaps between our KEMTL models and the baseline are larger in small dataset settings than in big

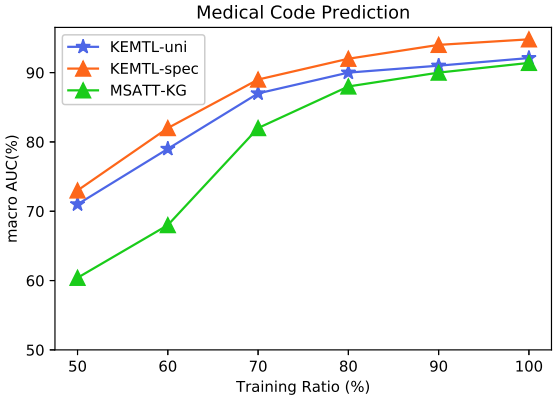

Figure 4: Model Performance Changing Curves with Different Training Size.

dataset settings. For example, with only 50% of the training data, the performance of our KEMTL-spec is still competitive, which shows an improvement of macro AUC 12.6% to MSATT-KG. We argue that this is because KEMTL can exploit the underlying inter-semantic and syntactic relationships that are inherently presented in different tasks, thus it can alleviate the data insufficiency problem and achieve good results with less data.

4.7 Case Study

In order to vividly show the benefits of the proposed multi-task learning framework, we visualize the detected informative concepts with high attention weights of a clinical text in the GAT-based encoding process. Here, we take an instance from "in hospital" mortality prediction task as an example in Figure 5. We display the results of three models: the single-task method KEMTL-Single, KEMTL-uni with universal sharing MTL and KEMTL-spec with task-specific sharing MTL. The colored part refers to extracted informative concepts. We can see from such visualization that for an eventually died patient, all of the three models can select the concepts about its personal information, common disease and symptoms. However, the KEMTL-single method fails to catch some symptoms such as "aspiration and atelectasis" and "pulmonary artery systolic hypertension", which may be not very common in this corpus. Nevertheless, as the KEMTL-uni and KEMTL-spec methods are able to take advantage of common knowledge from other tasks, it will be easier to capture these symptoms once they appear in other tasks and successfully make the right prediction. Compared with KEMTL-uni, KEMTL-spec can recognize more concepts describing the severity of physical condition and instrumental physical indicators like "severe" and "not well", which are very important for mortality prediction. This is because other than common knowledge, KEMTL-spec can also preserve the task specificity utilizing the task-specific encoder.

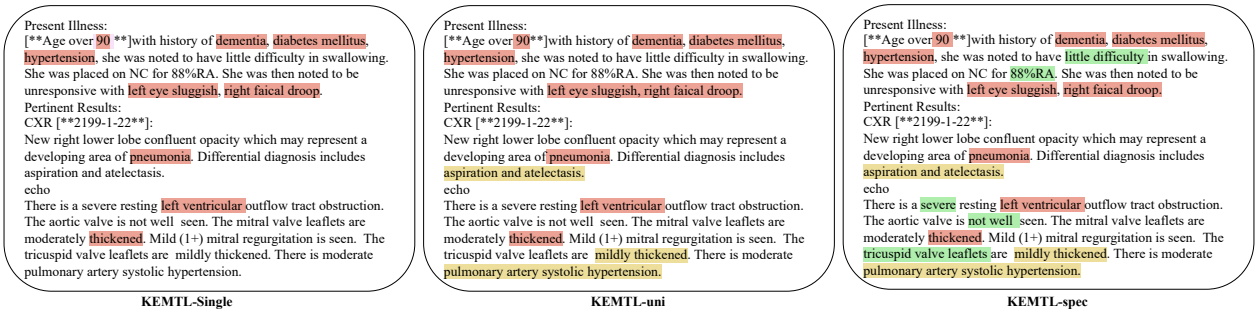

Figure 5: Visualization of Valued Concepts in Different Models.

5 RELATED WORK

5.1 Automatic ICD Coding

Automatic ICD coding is a hot research topic in the medical domain which has been studied since at least the 1990s [13]. Early approaches rely on human created features and utilize many machine learning models to capture them, such as SVM, KNN, Naive Bayes and topic model [2, 17, 21, 32, 35, 36]. Recent advanced approaches have employed deep neural networks for this task. Prakash et al. [33] utilized the memory network as well as Wikipedia to improve the accuracy of clinical diagnostic inferencing, which is the first study targeting at unstructured clinical texts for this task. Shi et al. [38] and Baumel et al. [4] utilized Recurrent Neural Network (RNN) with attention to generate hidden representations of written diagnosis descriptions and ICD codes. Besides, Convolutional Neural Networks (CNN) were explored for clinical note encoding and proven to be more effective than RNN-based models [23, 27, 29, 49]. Meanwhile, some works also proposed to utilize pre-training models for clinical texts encoding [1, 58]. Recently, to alleviate the long tail distribution of labels, several works tried to integrate external medical information into this task, such as the descriptions, co-occurrence, correlations and hierarchical dependency of ICD codes [6, 42, 44, 63]. Although the above methods are effective, they neglect the imbalance of the data itself while only dealing with the long-tail problem from the perspective of label space.

5.2 Graph Neural Networks (GNNs)

The concept of GNNs was first proposed in [34], which extended existing neural networks for processing the graph structured data. In recent years, GNNs have received growing attentions and variants of GNNs have been proposed, such as Graph Convolutional Network (GCN) [20] and Graph Attention Network (GAT) [43]. GNNs have also been exploring in several NLP tasks such as relation extraction [31], machine translation [28] and reading comprehension [9], where GNNs are used to encode semantics of texts. Recently there have been some studies exploring graph neural networks for general text classification. Peng et al. [30] converted each document into a graph based on word co-occurrence for large-scale hierarchical text classification. Yao et al. [53] and Yang et al. [51] followed this routine by taking advantage of the heterogeneous graph. These methods are focused on general text classification tasks rather than applications in the medical domain.

5.3 Multi-Task Learning (MTL)

MTL [8] is an approach to learn multiple tasks simultaneously, aiming at yielding performance gains from correlations and common features among related tasks. It has been widely adopted in different machine learning applications such as natural language processing [41] and computer vision [55]. Recently there have been some studies explored MTL in medical domains. For example, Chowdhury et al. [12] proposed an end-to-end multi-task encoder-decoder framework for three adverse drug reactions detection. Sun et al. [40] presented a multi-task aggregation network to share information across different coding schemes for medical code prediction. In this paper, we utilize the idea of MTL to take advantage of common information and knowledge for a variety of healthcare tasks.

6 CONCLUSION

In this paper, we devise a knowledge enhanced graph attention networks with multi-task learning for automatic ICD coding from clinical notes. Considering the data imbalance issue, we effectively utilize the concept-level and document-level correlations by generating a heterogeneous text graph to yield more semantically meaningful text representations. To future enable the cross-task information transition, we propose two auxiliary healthcare tasks: treatment recommendation and mortality prediction to boost the performance of ICD coding. In addition, by introducing medical domain knowledge, our model can alleviate the noise text issue more directly and effectively. Experimental results on the MIMIC-III medical dataset demonstrate that our proposed model outperforms state-of-the-art methods by a substantial margin.

ACKNOWLEDGEMENTS

This research was supported by National Social Science Fund of China (No. 22&ZD141), APRC - CityU New Research Initiatives (No.9610565, Start-up Grant for New Faculty of City University of Hong Kong), CityU - HKIDS Early Career Research Grant (No.9360163), Hong Kong ITC Innovation and Technology Fund Midstream Research Programme for Universities Project (No.ITS/034/22MS), SIRG - CityU Strategic Interdisciplinary Research Grant (No.7020046, No.7020074), SRG-Fd - CityU Strategic Research Grant (No.7005894), Tencent (CCF-Tencent Open Fund, Tencent Rhino-Bird Focused Research Fund), Huawei (Huawei Innovation Research Program), Ant Group (CCF-Ant Research Fund, Ant Group Research Fund) and Kuaishou.

REFERENCES

[1] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly Available Clinical BERT Embeddings. *CoRR* abs/1904.03323 (2019).

[2] Alan R. Aronson, Olivier Bodenreider, Dina Demner-Fushman, Kin Wah Fung, Vivian K. Lee, James G. Mork, Aurélie Névéol, Lee B. Peters, and Willie J. Rogers. 2007. From indexing the biomedical literature to coding clinical text: experience with MTI and machine learning approaches. In *BioNLP@ACL*. 105–112.

[3] Alan R. Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *J. Am. Medical Informatics Assoc.* 17, 3 (2010), 229–236.

[4] Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noemie Elhadad. 2018. Multi-label classification of patient notes: case study on ICD code assignment. In *AAAI Workshop*.

[5] Elena Birman-Deych, Amy D Waterman, Yan Yan, David S Nilasena, Martha J Radford, and Brian F Gage. 2005. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical care* (2005), 480–485.

[6] Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. HyperCore: Hyperbolic and Co-graph Representation for Automatic ICD Coding. In *ACL*. 3105–3114.

[7] Shilei Cao, Buyue Qian, Changchang Yin, Xiaoyu Li, Jishang Wei, Qinghua Zheng, and Ian Davidson. 2017. Knowledge Guided Short-Text Classification for Healthcare Applications. In *ICDM*. 31–40.

[8] Rich Caruana. 1997. Multitask Learning. *Machine Learning* 28, 1 (1997), 41–75.

[9] Kunlong Chen, Weidi Xu, Xingyi Cheng, Zou Xiaochuan, Yuyu Zhang, Le Song, Taifeng Wang, Yuan Qi, and Wei Chu. 2020. Question Directed Graph Attention Network for Numerical Reasoning over Text. In *EMNLP*. 6759–6768.

[10] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*. 1724–1734.

[11] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter F. Stewart. 2016. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In *NIPS*. 3504–3512.

[12] Shaika Chowdhury, Chenwei Zhang, and Philip S. Yu. 2018. Multi-Task Pharmacovigilance Mining from Social Media Posts. In *WWW*.

[13] Luciano R. S. de Lima, Alberto H. F. Laender, and Berthier A. Ribeiro-Neto. 1998. A Hierarchical Approach to the Automatic Categorization of Medical Documents. In *CIKM*. 132–139.

[14] Paulina Grnarova, Florian Schmidt, Stephanie L. Hyland, and Carsten Eickhoff. 2016. Neural Document Embeddings for Intensive Care Patient Mortality Prediction. *CoRR* abs/1612.00467 (2016).

[15] William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *NIPS*. 1024–1034.

[16] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3 (2016), 160035.

[17] Ramakanth Kavuluru, Anthony Rios, and Yuan Lu. 2015. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artif. Intell. Medicine* 65, 2 (2015), 155–166.

[18] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*. 1746–1751.

[19] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[20] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

[21] Bevan Koopman, Guido Zuccon, Anthony N. Nguyen, Anton Bergheim, and Narelle Grayson. 2015. Automatic ICD-10 classification of cancers from free-text death certificates. *Int. J. Medical Informatics* 84, 11 (2015), 956–965.

[22] Hung Le, Truyen Tran, and Svetha Venkatesh. 2018. Dual Control Memory Augmented Neural Networks for Treatment Recommendations. In *PAKDD*. 273–284.

[23] Fei Li and Hong Yu. 2020. ICD Coding from Clinical Text Using Multi-Filter Residual Convolutional Neural Network. In *AAAI*. 8180–8187.

[24] Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023. IMF: Interactive Multimodal Fusion Model for Link Prediction. In *WWW*. 2572–2580.

[25] Ning Liu, Pan Lu, Wei Zhang, and Jianyong Wang. 2019. Knowledge-Aware Deep Dual Networks for Text-Based Mortality Prediction. In *ICDE*. 1406–1417.

[26] Ziru Liu, Jiejie Tian, Qingpeng Cai, Xiangyu Zhao, Jingtong Gao, Shuchang Liu, Dayou Chen, Tonghao He, Dong Zheng, Peng Jiang, and Kun Gai. 2023. Multi-Task Recommendations with Reinforcement Learning. In *WWW*. 1273–1282.

[27] Junyu Luo, Cao Xiao, Lucas Glass, Jimeng Sun, and Fenglong Ma. 2021. Fusion: Towards Automated ICD Coding via Feature Compression. In *Findings of ACL*. 2096–2101.

[28] Diego Marcheggiani, Joost Bastings, and Ivan Titov. 2018. Exploiting Semantics in Neural Machine Translation with Graph Convolutional Networks. In *NAACL*.

[29] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *NAACL*. 1101–1111.

[30] Hao Peng, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang. 2018. Large-Scale Hierarchical Text Classification with Recursively Regularized Deep Graph-CNN. In *WWW*. 1063–1072.

[31] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. *TACL* 5 (2017), 101–115.

[32] Adler J. Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Gray Weiskopf, Frank D. Wood, and Noemie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *JAMIA* 21, 2 (2014), 231–237.

[33] Aaditya Prakash, Siyuan Zhao, Sadid A. Hasan, Vivek V. Datla, Kathy Lee, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2017. Condensed Memory Networks for Clinical Diagnostic Inferencing. In *AAAI*. 3274–3280.

[34] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The Graph Neural Network Model. *TNN* 20, 1 (2009), 61–80.

[35] Henning Schäfer and Christoph M. Friedrich. 2019. UMLS mapping and Word embeddings for ICD code assignment using the MIMIC-III intensive care database. In *EMBC*. 6089–6092.

[36] Elyne Scheurwegs, Boris Cule, Kim Luyckx, Léon Luyten, and Walter Daelemans. 2017. Selecting relevant features from the electronic health record for clinical code prediction. *J. Biomed. Informatics* 74 (2017), 92–103.

[37] Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019. GAMENet: Graph Augmented MEmory Networks for Recommending Medication Combination. In *AAAI*. 1126–1133.

[38] Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P. Xing. 2017. Towards Automated ICD Coding Using Deep Learning. *CoRR* abs/1711.04075 (2017).

[39] Congzheng Song, Shanghang Zhang, Najmeh Sadoughi, Pengtao Xie, and Eric P. Xing. 2020. Generalized Zero-Shot Text Classification for ICD Coding. In *IJCAI*. 4018–4024.

[40] Wei Sun, Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. 2021. Multitask Recalibrated Aggregation Network for Medical Code Prediction. *CoRR* abs/2104.00952 (2021).

[41] Bing Tian, Yong Zhang, Jin Wang, and Chunxiao Xing. 2019. Hierarchical Inter-Attention Network for Document Classification with Multi-Task Learning. In *IJCAI*. 3569–3575.

[42] Shang-Chi Tsai, Chao-Wei Huang, and Yun-Nung Chen. 2021. Modeling Diagnostic Label Correlation for Automatic ICD Coding. In *NAACL*. 4043–4052.

[43] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR*.

[44] Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A Label Attention Model for ICD Coding from Clinical Text. In *IJCAI*.

[45] Yejing Wang, Zhaocheng Du, Xiangyu Zhao, Bo Chen, Huifeng Guo, Ruiming Tang, and Zhenhua Dong. 2023. Single-shot Feature Selection for Multi-task Recommendations. In *SIGIR*. 341–351.

[46] Yejing Wang, Shen Ge, Xiangyu Zhao, Xian Wu, Tong Xu, Chen Ma, and Zhi Zheng. 2023. Doctor Specific Tag Recommendation for Online Medical Record Management. In *KDD*.

[47] Rui Wu, Zhaopeng Qiu, Jiacheng Jiang, Guilin Qi, and Xian Wu. 2022. Conditional Generation Net for Medication Recommendation. In *WWW*. 935–945.

[48] Cao Xiao, Edward Choi, and Jimeng Sun. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *JAMIA* 25, 10 (2018), 1419–1428.

[49] Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. 2019. EHR with Multi-scale Feature Attention and Structured Knowledge Graph Propagation. In *CIKM*. 649–658.

[50] Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021. SafeDrug: Dual Molecular Graph Encoders for Recommending Effective and Safe Drug Combinations. In *IJCAI*. 3735–3741.

[51] Tianchi Yang, Linmei Hu, Chuan Shi, Houye Ji, Xiaoli Li, and Liqiang Nie. 2021. HGAT: Heterogeneous Graph Attention Networks for Semi-supervised Short Text Classification. *TOIS* 39, 3 (2021), 32:1–32:29.

[52] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *NAACL*. 1480–1489.

[53] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph Convolutional Networks for Text Classification. In *AAAI*. 7370–7377.

[54] Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code Synonyms Do Matter: Multiple Synonyms Matching Network for Automatic ICD Coding. In *ACL*. 808–814.

[55] Tianzhu Zhang, Bernard Ghanem, Si Liu, and Narendra Ahuja. 2012. Robust visual tracking via multi-task sparse learning. In *CVPR*. 2042–2049.

[56] Yutao Zhang, Robert Chen, Jie Tang, Walter F. Stewart, and Jimeng Sun. 2017. LEAP: Learning to Prescribe Effective and Safe Treatment Combinations for Multimorbidity. In *SIGKDD*. 1315–1324.

[57] Yu Zhang and Qiang Yang. 2017. A Survey on Multi-Task Learning. *CoRR* abs/1707.08114 (2017).
[58] Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020. BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining. In *ClinicalNLP@EMNLP*. 24–34.
[59] Zijian Zhang, Xiangyu Zhao, Hao Miao, Chunxu Zhang, Hongwei Zhao, and Junbo Zhang. 2023. AutoSTL: Automated Spatio-Temporal Multi-Task Learning. In *AAAI*. 4902–4910.
[60] Zhi Zheng, Zhaopeng Qiu, Hui Xiong, Xian Wu, Tong Xu, Enhong Chen, and Xiangyu Zhao. 2022. DDR: Dialogue Based Doctor Recommendation for Online Medical Service. In *KDD*. 4592–4600.
[61] Zhi Zheng, Chao Wang, Tong Xu, Dazhong Shen, Penggang Qin, Baoxing Huai, Tongzhu Liu, and Enhong Chen. 2021. Drug Package Recommendation via Interaction-aware Graph Induction. In *WWW*. 1284–1295.
[62] Zhi Zheng, Chao Wang, Tong Xu, Dazhong Shen, Penggang Qin, Xiangyu Zhao, Baoxing Huai, Xian Wu, and Enhong Chen. 2023. Interaction-aware Drug Package Recommendation via Policy Gradient. *TOIS* 41, 1 (2023), 3:1–3:32.
[63] Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. 2021. Automatic ICD Coding via Interactive Shared Representation Networks with Self-distillation Mechanism. In *ACL*. 5948–5957.